Journal of Innovative Technology Convergence
*Vol. 2, No. 2 December 2020, pp. 105-112*

# Behavior Analysis of Convolutional Neural Network for Environmental Sound

**Ricardo A. Catanghal Jr.***

**Abstract:** Computer recognizing environmental sounds is a challenging and complex problem for a machine and an emerging field of research. In this study, the Convolutional Neural Network (CNN) behavior was analyzed against the environmental sounds. The performance level of the Convolutional Neural Network in identifying the environmental sounds using the parameters that we defined yields an excellent overall accuracy of 96.8%. This gives the model an excellent accurate prediction in identifying the given environmental sounds in the area of machine learning. The lowest accuracy among the group is the door knock, but the accuracy of 95.00%, still considered excellent currently in the field, and thus its parameters are fit for the environmental sounds.

**Keywords:** Convolutional Neural Network, Environmental Sound, Machine Learning, Acoustic

## 1. Introduction

Classification of environmental sounds plays a key role in security, investigation, robotics since the study of the sounds present in a specific environment can allow getting significant insights. The lack of standardized methods for an automatic and effective environmental sound classification (ESC) creates a need to be urgently satisfied. Environment sound is due to numerous sources present in the environment, such as living beings, non-living objects, and artificial entities created by humans. These sources contribute to the environment sound, which may be audible as well as non-audible to human ears. The sounds are captured mostly by acoustic sensors and radar systems [1][2] and subjected to further processing in various sound analysis and applications. The environment sounds, generated by various living beings and non-living objects, needs to be classified in certain categories in order to be used for different purposes, such as security, crime investigation, automated operation of robotic-like vehicles, weather forecasting, environment monitoring [3], and other applications. Current literature reports numerous studies and research contributions in the area of environment sound classification (ESC).

* College of Computer Studies, University of Antique, Sibalom, Antique, Philippines
Email: racatanghal@antiquespride.edu.ph

Acknowledgement: This paper has been supported and funded by the University of Antique.

*Received [September 12, 2020]; Revised [November 23, 2020]; Accepted [November 30, 2020]*



Published by InnoCon Publishing
ISSN 2704-4440

In general, most research on audio recognition has focused primarily on speech and music. The past researcher has taken this area of research in sound pattern recognition as the starting point because if we understood more about how humans hear, we could make machines hear better, in the sense of being able to analyze sound and extract useful and meaningful information from it [2]. In order to stimulate research in machine listening for general audio environments, in 2012–2013, IEEE Audio and Acoustic Signal Processing Technical Committee organized a research challenge on Detection and Classification of Acoustic Scenes and Events, that sounds from the environment or no-speech/non-music [4].

Environmental sound classification is an exciting field and is now a growing research problem in multimedia applications. The environmental sounds are a very diverse group of everyday audio events that cannot be described as only speech or music [5]. The environmental sounds are essential for understanding the content of the multimedia. Therefore, the environmental sound classification technology development is better for characterizing the essential role of environmental sounds in many smart homes technologies, such as home automation [6], audio surveillance system [7], hearing aids [8], smart room monitoring [9], and video content highlight generation [10], among others. Today, modern-day smart homes are now being equipped with artificial intelligence not just from the previous sentrollers (sensors, actuators, and controllers). The application of artificial intelligence is gaining momentum in automating a routinary task, and consumers and people agree that there is much potential usefulness in AI-run monitoring systems [11].

This study stems from the idea of extracting valuable information from the surroundings or environment in general and used in smart homes. It is clear that sounds carry a large amount of information about our everyday environment and physical events that take place in it, being able to detect which are the sound sources present in the signals would further increase the usefulness of any audio recording. The non-speech environmental sound framework was developed based on selected acoustic extracted features based on a sufficiently accurate classifier model, that was fine-tuned to get an improved parameter combination for the model to provide better classification and recognition.

## 2. Methodology

The first in the pipeline for acoustic recognition is the pre-processing, this is an integral step in Machine Learning as the quality of data, and the useful information that can be derived from it directly affects the ability of our model to learn [12]. This is a method that is used to convert the raw data into a clean data set, and it is extremely important that we preprocess our data before feeding it into our model. The three commonly sound-processing methods are presented: framing based, sub-framing, and sequential.

### 2.1 Signal Processing

The first in the pipeline Even what particular type of problem is applied in the acoustic field, the very basic structure of the system is the same and characterized using a universal or general pattern design shown in Figure 1.

Framing-based processing is a method wherein Audio signals to be classified are first divided into frames, often using a Hanning or a Hamming window. Features are extracted from each frame, and this set of features is used as one instance of training or testing. A classification decision is made for each frame and, hence, consecutive frames may belong to different classes. A major drawback of this processing scheme is that there is no way of selecting an optimal framing-window length suited for all classes. Some sound events are short-lived (*e.g.*, gun-shot) as compared with other longer events (e.g., thunder). If the window length is too small, then the long-term variations in the signal would not be well captured by the extracted features, and the framing method might chop events into multiple frames. On

the other hand, if the window length is too large, it becomes difficult to locate segmental boundaries between consecutive events, and there might be multiple sound events in a single frame. Also, one has to rely on features to extract nonstationary attributes of the signal since such a model does not allow the use of sequential learning methods [11][13].

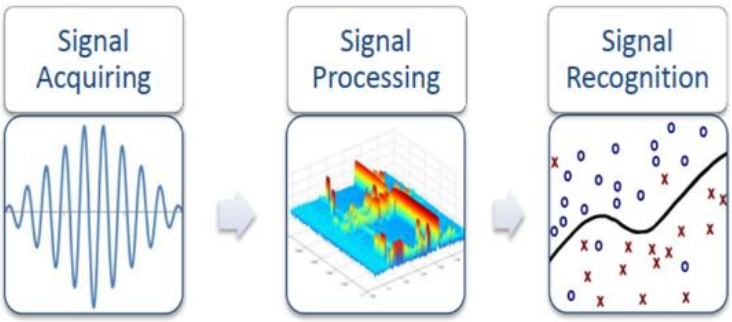


**Figure 1.** Signal Acquisition General Process

Audio feature extraction is a process that involves transforming audio data into a set of features such as pitch, timbre, and others. Specifically, the audio feature extraction process addresses the analysis and extraction of meaningful information from audio signals. The objective of the audio feature extraction process is to capture the relevant information on an audio signal to get a higher-level understanding of the audio signal. Furthermore, the extracted features of the audio signal may provide a higher-level understanding of the amplitude or frequency components of the audio signal by plotting to the spectrogram [14].

Figure 2 further depicts the detailed process in the conversion of acoustic features from the raw file. The first in the process is the acquisition of the signal: the acquisition of the continuous sound or audio stream is done through a device (The common is a microphone for example.) and dividing this into a block of the shorter signal called windowing.

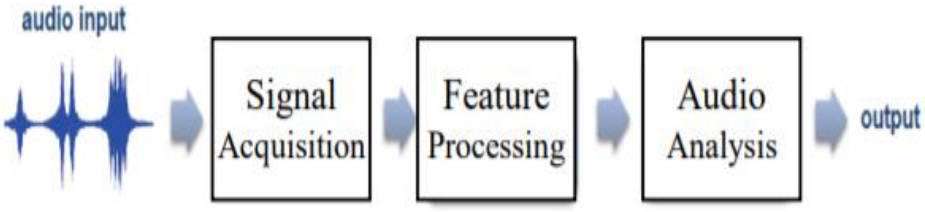


**Figure 2.** Detailed Steps

In order to carry out the windowing process, a theoretical continuous sample of sound streams from the input signal will be slid by a window function. Moreover, so that the examination of the ensuing signal be carried out, contingent on the window function duration it is presumed that a commonly non-static sound signal within the individual frame is quasi-stationary.

The issue of buffering is addressed by the implementation of the window function. The execution of the window function is performed by taking at a specified interval of time of audio chunks, even not knowing with regards to their completeness. In addressing the issue further with the use of the window function, the edge of the buffer is polished. With this, the entire essential period in the audio signal is required to be completed. In this paper, we applied a Hamming window and characterized it in equation one.

$$HammingWindow(s) = 0.54 - 0.46\cos\left(\frac{2\pi s}{BufferSize}\right) \quad (1)$$

### 2.2 Convolutional Neural Networks (ConvNets)

The Convolutional Neural Networks (ConvNets) are very much alike to the common Neural Networks. They are composed of neurons sometimes referred to as units or nodes that have learnable biases and weights. Although it has some practical effects which are consequential and meaningfully important because of its architectural design difference. Figure 3 shows a neural network with an input layer, convolutional layers (A mixture or combination can be performed in several ways.), layers that are completely linked and are hidden and limited in numbers and a layer for output (loss). This is the composition in an ordinary and conventional neural network in a deep architecture, a composition of a few distinct layers stacked collectively. While in comparison with the multilayer perceptron, the actual distinction lies in addition to the pooling and convolution procedures [8].

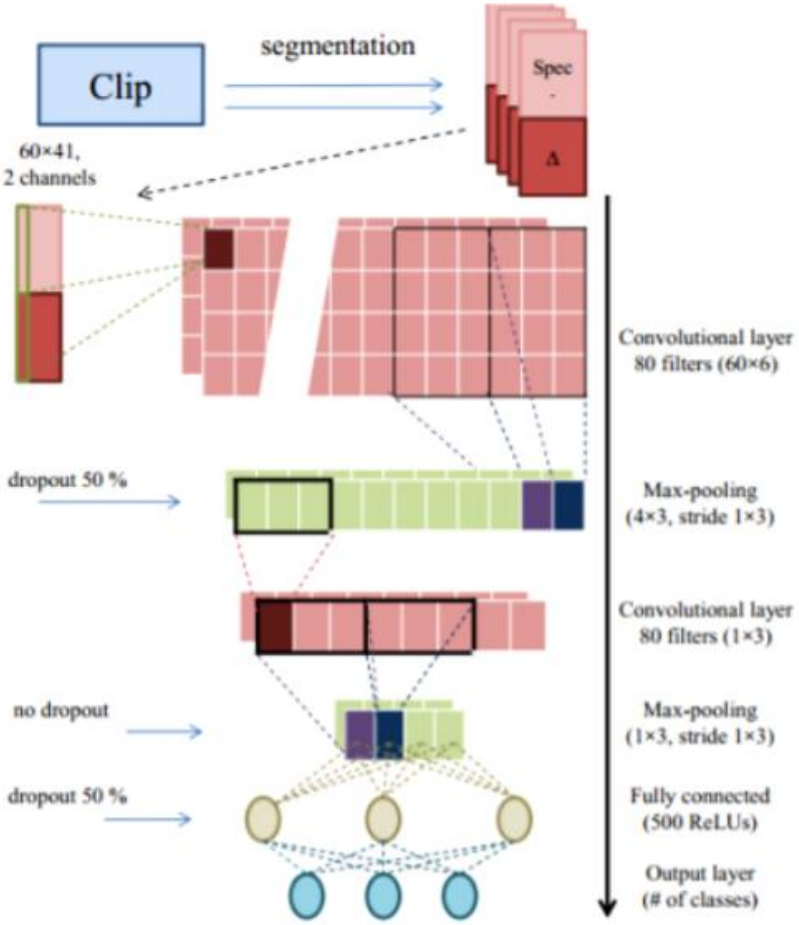


**Figure 3.** Convolutional Neural Network [5[11]

### 2.2 Sound Dataset

The sound datasets from this study were gathered by the researcher following the methods of previous work [15] with modification to fit in this study. Five sounds were chosen namely: Barking Dog, Crying Baby, Door Knock, Alarm Snoozing, and Glass Breaking. These five sounds play an essential role in smart homes, and as the study suggests. The consumers see that smart homes with an AI for

acoustics are of considerable value in providing them assistance in the specifics in the smart homes such as security, safety, child monitoring, assisted living, pet monitoring, and many others [15].

A 5-second-long recording of audio events (Shorter events were padded with silence as needed.). The labeled datasets were consequently arranged into five uniformly sized cross-validation folds, ensuring that clips originating from the same initial source file are always contained in a single fold.

## 3. Results and Discussion

Analyzing, the result of the essential tool for evaluating models, to have a better understanding of the performance [15]: the confusion matrix in Figure 4. The confusion matrix presents the summary of the prediction results and where it was confused when making a prediction. Let us start with the analysis on the "Baby Crying" it has an accuracy of 99.00%, the machine learning model that we have correctly labeled every "Baby Crying" as it is, with a confidence of 99.73%. Although some of the broken glass is classified as baby crying, bringing the precision to only 95%.

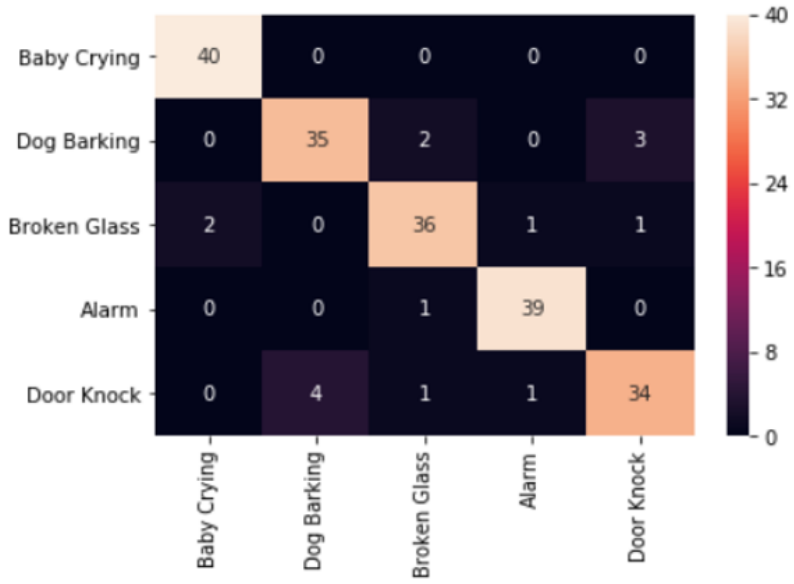


**Figure 4.** Confusion Matrix

The "Alarm Clock" which has the next highest true positive base on the confusion matrix has an accuracy of 98.50%. The only confusion the machine learning model for the alarm clock was with the broken glass. Further base on the result of the confusion matrix, the model confused broken glass and door knock as an alarm clock, bringing the Precision to 95.12%.

The "Broken Glass" has a true positive value of 36, which has an accuracy of 96.00% is considered still in a good performance. It can correctly identify 38 different broken glass sounds out of the forty samples. Although the model was confused with broken glass as baby crying, alarm, and door knock having a false positive rate of 2.50%, it still has a Precision of 90.00%. The machine learning model is confused with other sounds such as door knock, alarm clock, and dog barking as broken glass. In the confusion matrix out of the four confused, 50% is identified as the dog.

The dog barking has an accuracy of 95.50%, thirty-five out of 40 samples were correctly identified by the model. The model has a false positive rate of 2.50%, five out of forty samples were identified as different sound, broken glass, and door knock. Out of the five three or 60% is a door knock, which

implies that the model distinguishes door knock as a dog barking more likely when confused. The dog sound has 89.74% precision, which is acceptable.

## 4. Conclusion and Recommendations

Sounds carry a large amount of information about the everyday environment and physical events that take place in it. Developing signal processing methods to extract this information automatically has enormous potential in several applications, specifically smart homes in general. The performance level of the Convolutional Neural Network in identifying the environmental sounds using the parameters that we defined yields an excellent overall accuracy of 96.8%. This gives the model an excellent accurate prediction in identifying the given environmental sounds in the area of machine learning and is useful in different applications.

For future works, the following are recommended: integration of the model to the applications in order to be tested for its use. One of these is home security that is one of the current trends. The monitoring of homes or of the residents is one of the potentials among others. In the theoretical concept, the model needs to be further tested in different scenarios and situations, that is comparing the different performance and predictors in different situations. One of these is the comparison between internal and external settings.

## References

[1] P. Addabbo, M. di Bisceglie, C. Galdi, S. L. Ullo, "*The hyperspectral unmixing of trace-gases from ESA SCIAMACHY reflectance data*", IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 10, October 2015, pp.2130-2134, doi: 10.1109/LGRS.2015.2452315.

[2] R. Catanghal, T. Palaoag, C. Dayagdag, "*Environmental acoustic transformation and feature extraction for machine hearing*", in Proc. IOP Conference SerIes: Materials Science and Engineering, vol. 482, March 2019, pp.1-6, doi: 10.1088/1757-899X/482/1/012007.

[3] S. Ahmad, S. Agrawal, S. Joshi, S. Taran, V. Bajaj, F. Demir, A. Sengur, "*Environmental sound classification using optimum allocation sampling based empirical mode decomposition*", Physica A: Statistical Mechanics and its Applications, vol. 537, January 2020, doi: 10.1016/j.physa.2019.122613.

[4] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, M. D. Plumbley, "*Detection and classification of acoustic scenes and events*", IEEE Transactions on Multimedia, vol. 17, no. 10, October 2015, pp.1733–1746, doi: 10.1109/TMM.2015.2428998.

[5] D. M. Agrawal, H. B. Sailor, M. H. Soni, H. A. Patil, "*Novel TEO-based gammatone features for environmental sound classification*", 2017 25th European Signal Processing Conference (EUSIPCO), Kos, Greece, August 28-September 2, 2017, pp.1809-1813, doi: 10.23919/EUSIPCO.2017.8081521.

[6] N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, H. G. Okuno, "*Environmental sound recognition for robot audition using matching-pursuit*", in *Modern Approaches in Applied Intelligence*, K.G. Mehrotra, C.K. Mohan, J.C. Oh, P.K. Varshney, M. Ali, Eds., Lecture Notes in Computer Science, vol. 6704. Springer, Berlin, Heidelberg, doi: 10.1007/978-3-642-21827-9_1.

[7] F. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, M. Vento, "*Audio surveillance of roads: a system for detecting anomalous sounds*", IEEE TransActions on Intelligent Transportation Systems, vol. 17, no. 1, January 2016, pp. 279–288, doi: 10.1109/TITS.2015.2470216.

[8] E. Alexandre, L. Cuadra, M. Rosa, F. Lopez-Ferreras, "*Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms*", IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 8, November 2007, pp.2249–2256, doi: 10.1109/TASL.2007.905139.

[9] M. Vacher, J. -F. Serignat, S. Chaillol, "*Sound classification in a smart room environment: an approach using GMM and HMM methods*", in The 4th IEEE Conference on Speech Technology and Human-Computer Dialogue (SpeD), vol. 1, May 2007, pp.135–146.

[10] L. Ballan, A. Bazzica, M. Bertini, A. Del Bimbo, G. Serra, "*Deep networks for audio event classification in soccer videos*" 2009 IEEE International Conference on Multimedia and Expo, New York, USA, June 28-July 3, 2009, pp.474–477, doi: 10.1109/ICME.2009.5202537.

[11] R. A. Catanghal, T. D. Palaoag, C. Dayagdag, "*Meta-Analysis of Acoustic Feature Extraction for Machine Listening Systems*", in Proc. of the 2019 8th International Conference on Software and Computer Applications, February 2019, pp.369-372, doi: 10.1145/3316615.3316664.

[12] D. Kumar, "*Introduction to Data Preprocessing in Machine Learning: Beginners Guide for Data Preprocessing*", towardsdatascience.com, www.towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d (accessed August 9, 2020).

[13] S. Chachada, C. C. J. Kuo, "*Environmental sound recognition: A survey*", 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Kaohsiung, Taiwan, October 29-November 1, 2013, doi: 10.1109/APSIPA.2013.6694338.

[14] N. Dave, "*Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition*", International Journal of Research in Engineering and Advanced Technology, vol. 1, no. 6, July 2013, pp.1-5.

[15] R. Catanghal, "*A Framework for Home Machine Listening System with Convolutional Neural Network*", International journal of simulation: systems, science & technology, 2019, doi: 10.5013/ijssst.a.20.s2.08.